Санкт-Петербургский Политехнический Университет им. Петра Великого

Институт прикладной математики и механики

Кафедра прикладной математики

**Отчёт по лабораторным работам №5-8 по дисциплине “Математическая статистика”**

Выполнил студент:

Мишутин Д. В.

Группа:

3630102/70301

Проверил:

К.ф.-м.н., доцент

Баженов Александр Николаевич

Санкт-Петербург

2020 г.

Оглавление

[1 Постановка задачи 4](#_Toc42888603)

[1.1 Выборочные коэффициенты корреляции и эллипсы рассеивания 4](#_Toc42888604)

[1.2 Простая линейная регрессия 4](#_Toc42888605)

[1.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат 4](#_Toc42888606)

[1.4 Интервальные оценки математического ожидания и стандартного отклонения 4](#_Toc42888607)

[2 Теория 5](#_Toc42888608)

[2.1 Выборочные коэффициенты корреляции и эллипсы рассеивания 5](#_Toc42888609)

[2.2 Простая линейная регрессия 5](#_Toc42888610)

[2.2.1 Метод наименьших квадратов (МНК) 5](#_Toc42888611)

[2.2.2 Метод наименьших модулей (МНМ) 6](#_Toc42888612)

[2.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат 6](#_Toc42888613)

[2.3.1 Метод максимального правдоподобия (ММП) 6](#_Toc42888614)

[2.3.2 Критерий Пирсона 7](#_Toc42888615)

[2.4 Интервальные оценки математического ожидания и стандартного отклонения 7](#_Toc42888616)

[2.4.1 Интервальные оценки 8](#_Toc42888617)

[2.4.2 Асимптотические оценки 8](#_Toc42888618)

[3 Реализация 8](#_Toc42888619)

[4 Результаты 9](#_Toc42888620)

[4.1 Выборочные коэффициенты корреляции и эллипсы рассеивания 9](#_Toc42888621)

[4.1.1 Таблицы 9](#_Toc42888622)

[4.1.2 Иллюстрации 11](#_Toc42888623)

[4.2 Простая линейная регрессия 19](#_Toc42888624)

[4.2.1 Таблицы 19](#_Toc42888625)

[4.2.2 Иллюстрации 20](#_Toc42888626)

[4.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат 20](#_Toc42888627)

[4.3.1 Нормальное распределение 20](#_Toc42888628)

[4.3.2 Равномерное распределение 21](#_Toc42888629)

[4.4 Интервальные оценки математического ожидания и стандартного отклонения 21](#_Toc42888630)

[5 Выводы 21](#_Toc42888631)

[5.1 Выборочные коэффициенты корреляции и эллипсы рассеивания 21](#_Toc42888632)

[5.2 Простая линейная регрессия 22](#_Toc42888633)

[5.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат 22](#_Toc42888634)

[5.4 Интервальные оценки математического ожидания и стандартного отклонения 22](#_Toc42888635)

[6 Литература 22](#_Toc42888636)

[7 Приложения 22](#_Toc42888637)

Список иллюстраций и таблиц

[Выборочные коэффициенты корреляции и эллипсы рассеивания](#лаба_5)*………………………………….……………….…9*

[Простая линейная регрессия](#лаба_6)*…………………………………………………………………………………………………………….…19*

[Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат](#лаба_7)………..20

[Интервальные оценки математического ожидания и стандартного отклонения](#лаба_8)………………………….…21

# 1 Постановка задачи

## Выборочные коэффициенты корреляции и эллипсы рассеивания

Сгенерировать двумерные выборки размерами 20, 60, 100 для нормального двумерного распределения . Коэффициент корреляции взять равным 0, 0.5, 0.9. Каждая выборка генерируется 1000 раз и для неё вычисляются: среднее значение, среднее значение квадрата и дисперсия коэффициентов корреляции Пирсона, Спирмена и квадрантного коэффициента корреляции. Повторить все вычисления для смеси нормальных распределений:

Изобразить сгенерированные точки на плоскости и нарисовать эллипс рассеяния.

## Простая линейная регрессия

Найти оценки коэффициентов линейной регрессии , используя 20 точек на отрезке с равномерным шагом равным 0.2. Ошибку считать нормально распределённой с параметрами . В качестве эталонной зависимости взять . При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей.

Проделать то же самое для выборки, у которой в значения и вносятся возмущения 10 и -10.

## Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Сгенерировать выборку объёмом 100 элементов для стандартного нормального распределения . По сгенерированной выборке оценить параметры и нормального закона методом максимального правдоподобия. В качестве основной гипотезы будем считать, что сгенерированное распределение имеет вид . Проверить основную гипотезу, используя критерий согласия . В качестве уровня значимости взять . Привести таблицу вычислений .

## 1.4 Интервальные оценки математического ожидания и стандартного отклонения

Для двух выборок из 20 и 100 элементов, сгенерированных согласно нормальному закону , для параметров масштаба и положения построить асимптотически нормальные интервальные оценки на основе точечных оценок метода максимального правдоподобия и классические интервальные оценки на основе статистик и Стьюдента. В качестве параметра надёжности взять .

# **Теория**

## Выборочные коэффициенты корреляции и эллипсы рассеивания

1. Двумерное стандартное нормальное распределение:
2. Коэффициент корреляции Пирсона:
3. Коэффициент корреляции Спирмена:
4. Квадрантный коэффициент корреляции:

## 2.2 Простая линейная регрессия

Стандартное нормальное распределение:

Точная формула:

где – заданные числа, – наблюдаемые значения, – независимые, одинаково распределённые значения ошибок с параметрами и , и – неизвестные параметры, подлежащие оценке.

### 2.2.1 Метод наименьших квадратов (МНК)

Критерий – минимизация RSS-функции (***R****esidual* ***S****um of* ***S****quares*):

В случае линейной регрессии оценочные коэффициенты и можно вычислить по формулам:

МНК является несмещённой оценкой. Чувствителен к выбросам (т. к. в вычислениях используется выборочное среднее, крайне неустойчивое к редким, но больших по величине выбросам).

### 2.2.2 Метод наименьших модулей (МНМ)

Критерий – минимизация LAD-функции (***L****east* ***A****bsolute* ***D****eviations*):

Коэффициенты так же можно вычислить по формулам:

МНМ-оценки обладают свойством робастности. Но на практике решение реализуется только численно.

## Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Нормальное распределение:

### Метод максимального правдоподобия (ММП)

МНМ – метод оценивания неизвестного параметра путём максимизации функции правдоподобия :

Оценкой максимального правдоподобия будем называть такое значение из множества допустимых значений , для которого принимает максимальное значение для заданных .

Тогда при оценивании математического ожидания и дисперсии нормального распределения получим:

Отсюда находятся выражения для оценок и :

### Критерий Пирсона

Разобьём генеральную совокупность на непересекающихся подмножеств , где – вероятность того, что точка попала в -ый промежуток.

Так как генеральная совокупность это , то крайние промежутки будут бесконечными:

Пусть – частота попадания выборочных элементов в .

В случае справедливости гипотезы относительно частоты при больших должны быть близки к , значит в качестве меры имеет смысл взять:

Тогда

Для выполнения гипотезы должны выполняться следующие условия:

где – квантиль распределения с степенями свободы порядка , – заданный уровень значимости.

## 2.4 Интервальные оценки математического ожидания и стандартного отклонения

Стандартное нормальное распределение:

Функция распределения Стьюдента:

Функция плотности :

Доверительным интервалом или интервальной оценкой числовой характеристики или параметра распределения с доверительной вероятностью называется интервал со случайными границами , содержащий с вероятностью .

### 2.4.1 Интервальные оценки

Интервальные оценки для математического ожидания нормального распределения:

где – квантиль распределения Стьюдента порядка .

Интервальные оценки для стандартного отклонения нормального распределения:

где и – квантили распределения Стьюдента порядков и соответственно.

### 2.4.2 Асимптотические оценки

Асимптотическая интервальная оценка для произвольного распределения при большой выборке математического ожидания:

Асимптотическая интервальная оценка для произвольного распределения при большой выборке стандартного отклонения:

где – квантиль стандартного нормального распределения порядка , , .

# 3 Реализация

Был использован язык *Python 3.8.2*: модуль *numpy* для вычисления описательных статистик, модуль *scipy* для генерации выборок и расчёта коэффициентов, модуль *matplotlib* для построения и отображения графиков, модуль *pandas* для хранения статистических данных в таблицах и функция *display* из модуля *IPython.display* для их корректного отображения.

# 4 Результаты

## Выборочные коэффициенты корреляции и эллипсы рассеивания

### 4.1.1 Таблицы

Таблица 1 Двумерное стандартное нормальное распределение, n=20, r=0

|  |  |  |  |
| --- | --- | --- | --- |
| **n=20** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.012 | 0.011 | 0.003 |
| **E(z^2)** | 0.048 | 0.05 | 0.051 |
| **D(z)** | 0.048 | 0.049 | 0.051 |

Таблица 2 Двумерное стандартное нормальное распределение, n=60, r=0

|  |  |  |  |
| --- | --- | --- | --- |
| **n=60** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.004 | 0.004 | 0.003 |
| **E(z^2)** | 0.016 | 0.017 | 0.017 |
| **D(z)** | 0.016 | 0.017 | 0.017 |

Таблица 3 Двумерное стандартное нормальное распределение, n=100, r=0

|  |  |  |  |
| --- | --- | --- | --- |
| **n=100** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.001 | 0.001 | -0.001 |
| **E(z^2)** | 0.01 | 0.01 | 0.011 |
| **D(z)** | 0.01 | 0.01 | 0.011 |

Таблица 4 Двумерное стандартное нормальное распределение, n=20, r=0.5

|  |  |  |  |
| --- | --- | --- | --- |
| **n=20** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.489 | 0.46 | 0.322 |
| **E(z^2)** | 0.271 | 0.246 | 0.148 |
| **D(z)** | 0.032 | 0.035 | 0.044 |

Таблица 5 Двумерное стандартное нормальное распределение, n=60, r=0.5

|  |  |  |  |
| --- | --- | --- | --- |
| **n=60** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.496 | 0.476 | 0.331 |
| **E(z^2)** | 0.255 | 0.237 | 0.124 |
| **D(z)** | 0.009 | 0.01 | 0.014 |

Таблица 6 Двумерное стандартное нормальное распределение, n=100, r=0.5

|  |  |  |  |
| --- | --- | --- | --- |
| **n=100** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.497 | 0.477 | 0.331 |
| **E(z^2)** | 0.253 | 0.233 | 0.118 |
| **D(z)** | 0.005 | 0.006 | 0.009 |

Таблица 7 Двумерное стандартное нормальное распределение, n=20, r=0.9

|  |  |  |  |
| --- | --- | --- | --- |
| **n=20** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.896 | 0.867 | 0.696 |
| **E(z^2)** | 0.806 | 0.756 | 0.513 |
| **D(z)** | 0.002 | 0.004 | 0.029 |

Таблица 8 Двумерное стандартное нормальное распределение, n=60, r=0.9

|  |  |  |  |
| --- | --- | --- | --- |
| **n=60** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.898 | 0.883 | 0.707 |
| **E(z^2)** | 0.808 | 0.78 | 0.508 |
| **D(z)** | 0.001 | 0.001 | 0.009 |

Таблица 9 Двумерное стандартное нормальное распределение, n=100, r=0.9

|  |  |  |  |
| --- | --- | --- | --- |
| **n=100** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | 0.899 | 0.886 | 0.708 |
| **E(z^2)** | 0.809 | 0.786 | 0.507 |
| **D(z)** | 0 | 0.001 | 0.005 |

Таблица 10 Смесь распределений, n=20

|  |  |  |  |
| --- | --- | --- | --- |
| **n=20** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | -0.08 | -0.078 | -0.05 |
| **E(z^2)** | 0.061 | 0.061 | 0.056 |
| **D(z)** | 0.054 | 0.055 | 0.054 |

Таблица 11 Смесь распределений, n=60

|  |  |  |  |
| --- | --- | --- | --- |
| **n=60** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | -0.092 | -0.086 | -0.06 |
| **E(z^2)** | 0.025 | 0.024 | 0.021 |
| **D(z)** | 0.016 | 0.016 | 0.017 |

Таблица 12 Смесь распределений, n=100

|  |  |  |  |
| --- | --- | --- | --- |
| **n=100** | **Pearson** | **Spearman** | **quadrant** |
| **E(z)** | -0.097 | -0.092 | -0.063 |
| **E(z^2)** | 0.019 | 0.019 | 0.014 |
| **D(z)** | 0.01 | 0.01 | 0.01 |

### 4.1.2 Иллюстрации

Рис. 1 Двумерное стандартное нормальное распределение для n=20, r=0Изображение выглядит как внутренний, текст, карта, фотография

Автоматически созданное описание

Рис. 2 Двумерное стандартное нормальное распределение для n=60, r=0Изображение выглядит как фотография, стол, компьютер, сидит

Автоматически созданное описание

Рис. 3 Двумерное стандартное нормальное распределение для n=100, r=0Изображение выглядит как фотография, компьютер, стол, белый

Автоматически созданное описание

Рис. 4 Двумерное стандартное нормальное распределение для n=20, r=0.5Изображение выглядит как карта, текст

Автоматически созданное описание

Рис. 5 Двумерное стандартное нормальное распределение для n=60, r=0.5Изображение выглядит как фотография, стол, полный, белый

Автоматически созданное описание

Рис. 6 Двумерное стандартное нормальное распределение для n=100, r=0.5

Изображение выглядит как фотография, стол, другой, компьютер

Автоматически созданное описание

Рис. 7 Двумерное стандартное нормальное распределение для n=20, r=0.9Изображение выглядит как фотография, мужчина, другой, стол

Автоматически созданное описание

Рис. 8 Двумерное стандартное нормальное распределение для n=60, r=0.9Изображение выглядит как фотография, стол, мужчина, снег

Автоматически созданное описание

Рис. 9 Двумерное стандартное нормальное распределение для n=100, r=0.9Изображение выглядит как фотография, мужчина, стол, снег

Автоматически созданное описание

Рис. 10 Смесь распределений для n=20Изображение выглядит как карта

Автоматически созданное описание

Рис. 11 Смесь распределений для n=60Изображение выглядит как фотография, стол, белый, компьютер

Автоматически созданное описание

Рис. 12 Смесь распределений для n=100

Изображение выглядит как внутренний, фотография, компьютер, стол

Автоматически созданное описание

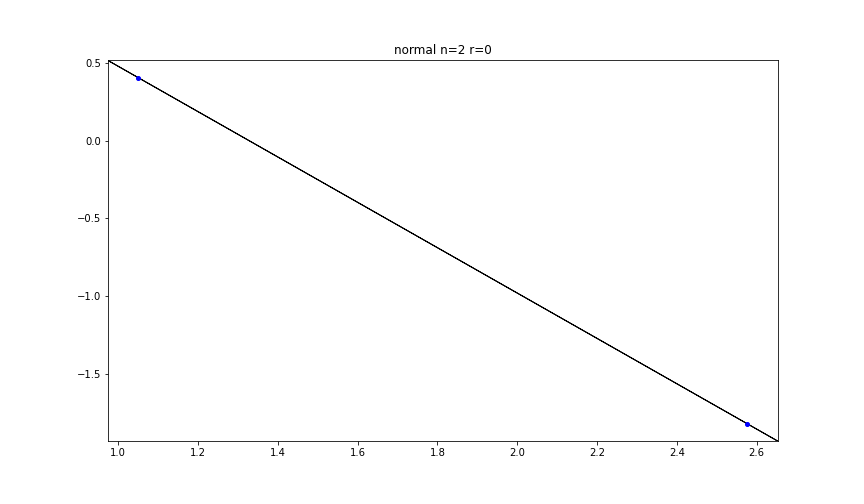
Рис. 13 Эллипс рассеивания для 2-х точек при r=0

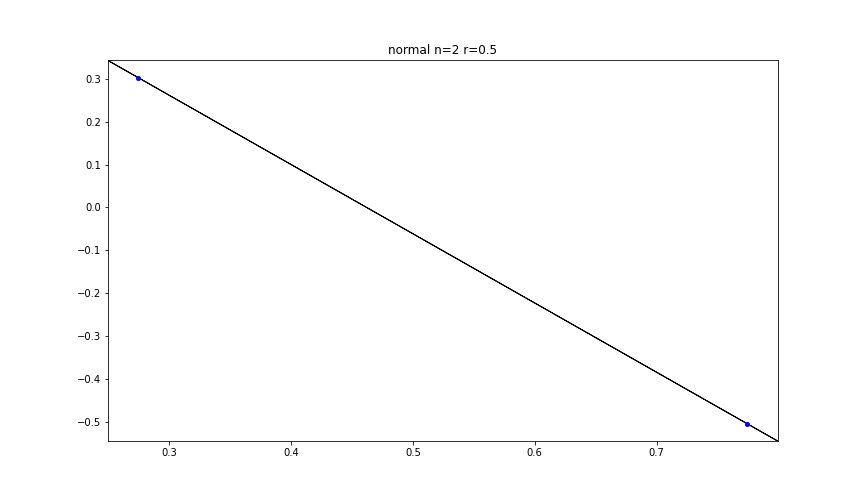
Рис. 14 Эллипс рассеивания для 2-х точек при r=0.5

Рис. 15 Эллипс рассеивания для 2-х точек при r=0.9Изображение выглядит как снимок экрана

Автоматически созданное описание

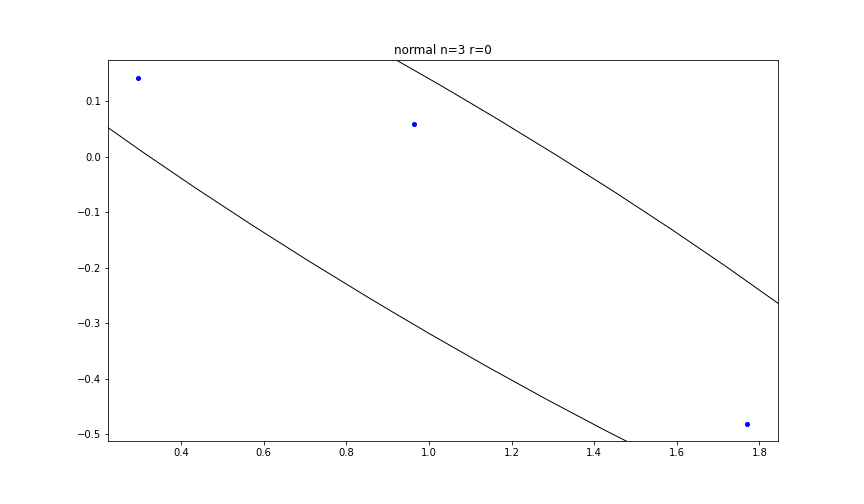
Рис. 16 Эллипс рассеивания для 3-х точек при r=0

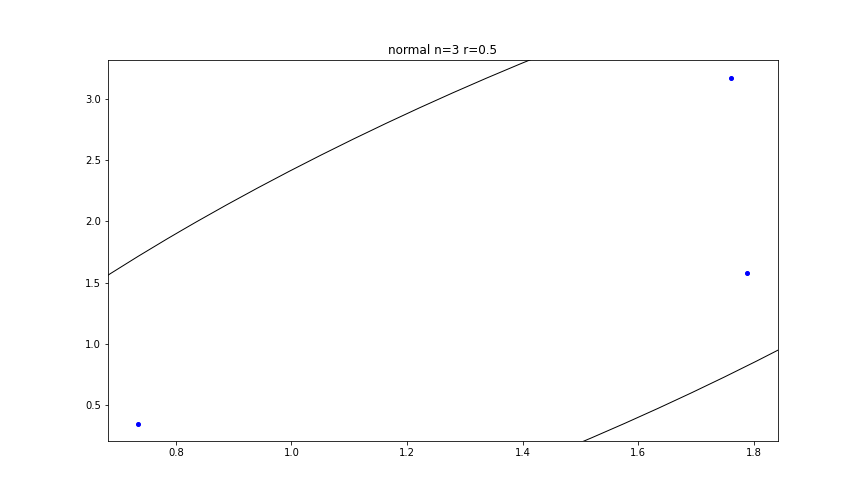
Рис. 17 Эллипс рассеивания для 3-х точек при r=0.5

Рис. 18 Эллипс рассеивания для 3-х точек при r=0.9

Изображение выглядит как снимок экрана

Автоматически созданное описание

## Простая линейная регрессия

### Таблицы

Таблица 1 Коэффициенты при выборке без возмущений

|  |  |  |
| --- | --- | --- |
| Исходная выборка |  |  |
| RSS | 1.70073 | 2.599262 |
| LAD | 1.75065 | 2.631271 |

Таблица 2 Коэффициенты при выборке с возмущениями

|  |  |  |
| --- | --- | --- |
| Выборка с возмущениями |  |  |
| RSS | 0.272159 | 2.742119 |
| LAD | 1.751878 | 2.631394 |

### Иллюстрации

Рис. 1 График получившейся линейной регрессии

Изображение выглядит как карта, текст

Автоматически созданное описание

## Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

### 4.3.1 Нормальное распределение

При подсчёте оценок параметров закона нормального распределения с помощью МНМ были получены следующие результаты:

Таблица 2 Вычисления

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **1** | -1 | 15.0 | 0.1465 | 0.0083 |
| **2** | -0.5 | 14.0 | 0.1428 | 0.0055 |
| **3** | 0 | 16.0 | 0.1870 | 0.3909 |
| **4** | 0.5 | 27.0 | 0.1925 | 3.1189 |
| **5** | 1 | 10.0 | 0.1557 | 1.9923 |
| **6** | inf | 18.0 | 0.1755 | 0.0118 |

### 4.3.2 Равномерное распределение

Размер выборки , заданный отрезок .

Таблица 3 Вычисления

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **1** | -2 | 0.0 | 0.0120 | 0.2399 |
| **2** | 1 | 18.0 | 0.8693 | 0.0216 |
| **3** | 4 | 2.0 | 0.1187 | 0.0587 |
| **4** | inf | 0.0 | 0.0000 | 0.0000 |

## 4.4 Интервальные оценки математического ожидания и стандартного отклонения

Таблица 1 Результаты для выборок мощности n=20

|  |  |  |
| --- | --- | --- |
| **n=20** |  |  |
| **normal\_dist** |  |  |
| **random\_dist** |  |  |

Таблица 2 Результаты для выборок мощности n=100

|  |  |  |
| --- | --- | --- |
| **n=100** |  |  |
| **normal\_dist** |  |  |
| **random\_dist** |  |  |

# 5 Выводы

## 5.1 Выборочные коэффициенты корреляции и эллипсы рассеивания

Ближе всего к теоретическому коэффициенту корреляции находится коэффициент Пирсона.

По графикам видно, что

* при увеличении объёма выборки коэффициенты корреляции стремятся к теоретическим
* при уменьшении корреляции эллипс рассеивания стремится к окружности, а при увеличении – вырождается в прямую с углом наклона в против часовой стрелки
* для построения эллипса рассеивания нужно минимум 3 точки, а при 2-х точках эллипс вырождается в прямую под определённым углом

## 5.2 Простая линейная регрессия

По графикам видно, что оба метода дают хорошую оценку, если нет выбросов. Однако выбросы сильно влияют на оценки по МНК.

Выбросы слабо влияют на оценку по МНМ, но ценой за это является бо́льшая вычислительная сложность.

## 5.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Табличное значение квантиля . Полученное значение критерия согласия Пирсона для нормального распределения следовательно основная гипотеза не может быть опровергнута на уровне значимости .

Для равномерного распределения полученное значение критерия Пирсона означает, что из полученной выборки мы не можем опровергнуть гипотезу о нормальности данного распределения.

## 5.4 Интервальные оценки математического ожидания и стандартного отклонения

Точность оценок растёт с увеличением объёма выборки, оба метода показывают примерно одинаковое качество оценок, но у асимптотического подхода (**random\_dist**) очевидное преимущество.

# 6 Литература

[Основы работы с *numpy* (отдельная глава курса)](https://stepik.org/course/401)

[Pandas обзор](https://www.dataquest.io/blog/pandas-cheat-sheet/)

[Документация по *scipy*](https://docs.scipy.org/doc/scipy/reference/index.html)

[Таблица значений](http://statsoft.ru/home/textbook/modules/sttable.html#chi)

# 7 Приложения

[Код лабораторной](https://github.com/MeShootIn/matstat/blob/master/lab_5/lab_5.ipynb) №5

[Код лабораторной](https://github.com/MeShootIn/matstat/blob/master/lab_6/lab_6.ipynb) №6

[Код лабораторной](https://github.com/MeShootIn/matstat/blob/master/lab_7/lab_7.ipynb) №7

[Код лабораторной](https://github.com/MeShootIn/matstat/blob/master/lab_8/lab_8.ipynb) №8